Chapter 6

Foundations of Business Intelligence: Databases and Information Management

Learning Objectives

- What are the problems of managing data resources in a traditional file environment?
- What are the major capabilities of database management systems (DBMS) and why is a relational DBMS so powerful?
- What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?
- Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?

Better Data Management Helps Toronto Globe and Mail

- **Problem:**
  - Data fragmented in isolated databases and files
  - Time-consuming reporting processes
  - Outdated data management technology
- **Solution:**
  - Replace disparate systems with enterprise system, with centralized mainframe and data management system

Better Data Management Helps Toronto Globe and Mail

- SAP enterprise system with SAP NetWeaver BW data warehouse to contain all company's data; educate users and tools
- Demonstrates IT's role in successful data management
- Illustrates digital technology's ability to lower costs while improving performance

- **File organization concepts**
  - **Database: Group of related files**
  - **File: Group of records of same type**
  - **Record: Group of related fields**
  - **Field: Group of characters as word(s) or number**
    - Describes an **entity** (person, place, thing on which we store information)
    - **Attribute:** Each characteristic, or quality, describing entity
      - Example: Attributes DATE or GRADE belong to entity COURSE

THE DATA HIERARCHY

A computer system organizes data in a hierarchy that starts with the bit, which represents either a 0 or a 1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can be collected to form a file, and related files can be organized into a database.

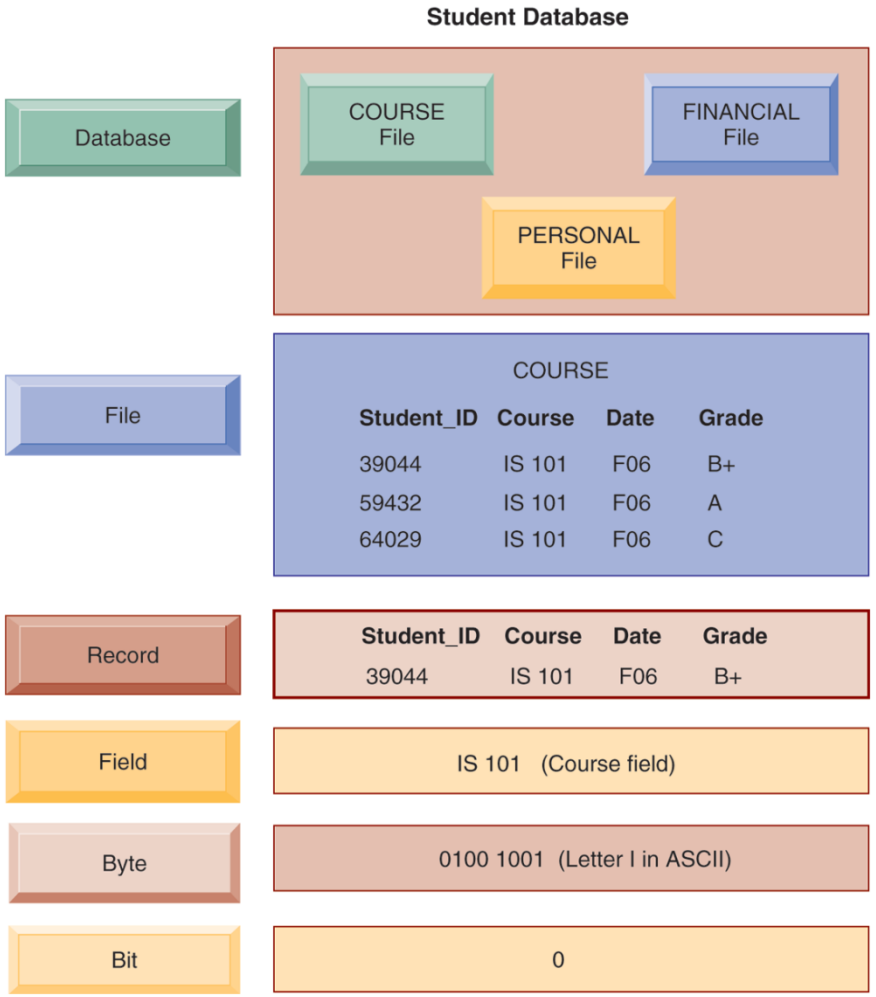


FIGURE 6-1

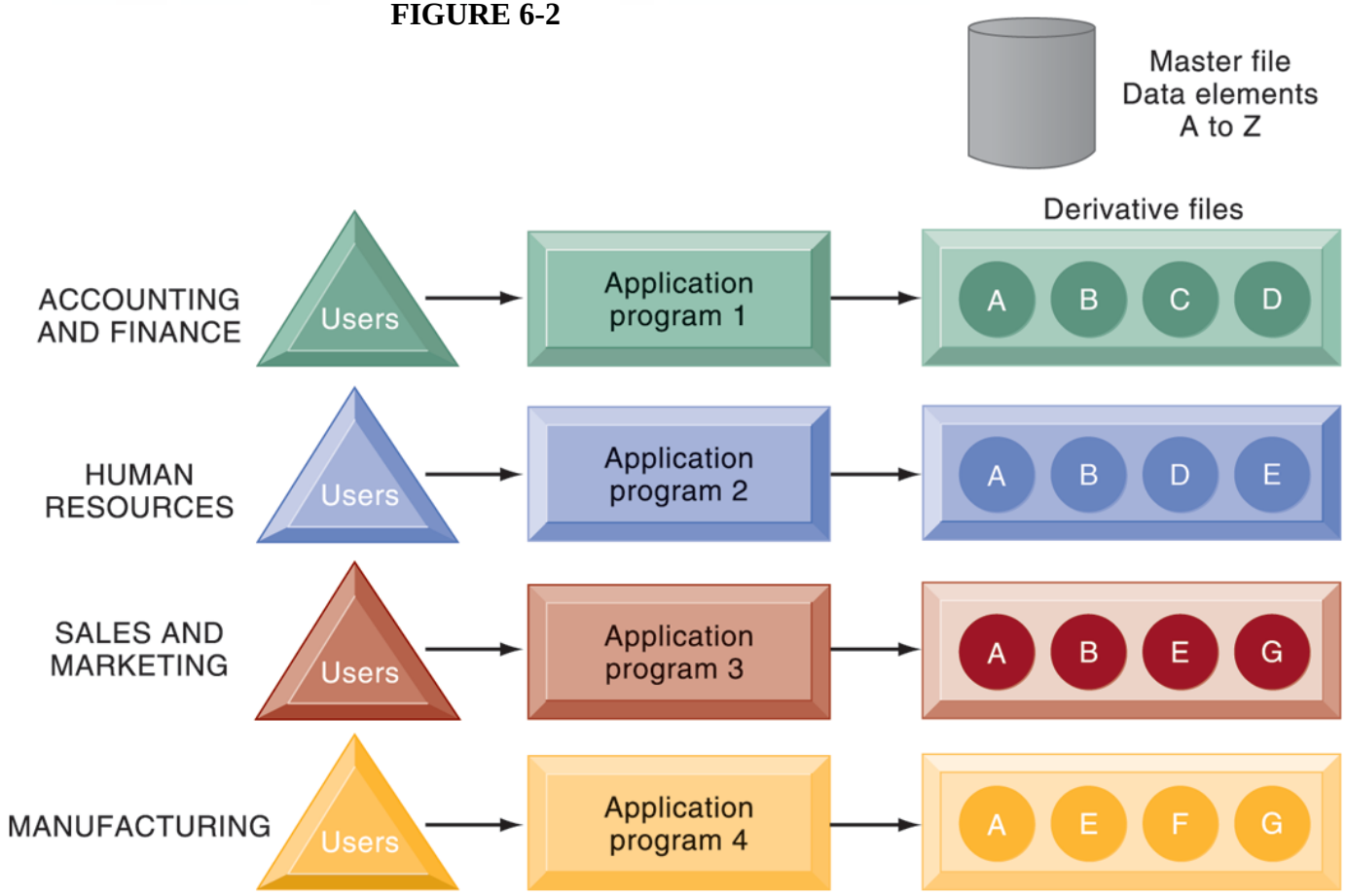
Managing Data in a Traditional File Environment

- **Problems with the traditional file environment (files maintained separately by different departments)**
  - **Data redundancy:**
    - Presence of duplicate data in multiple files
  - **Data inconsistency:**
    - Same attribute has different values
  - **Program-data dependence:**
    - When changes in program requires changes to data accessed by program
  - **Lack of flexibility**
  - **Poor security**
  - **Lack of data sharing and availability**

TRADITIONAL FILE PROCESSING

The use of a traditional approach to file processing encourages each functional area in a corporation to develop specialized applications. Each application requires a unique data file that is likely to be a subset of **the master file. These subsets of the master file lead to data redundancy and inconsistency, processing inflexibility, and wasted storage** resources.

FIGURE 6-2



Capabilities of Database Management Systems (DBMSs)

- **Database**
  - **Serves many applications by centralizing data and controlling redundant data**
- **Database management system (DBMS)**
  - **Interfaces between applications and physical data files**
  - **Separates logical and physical views of data**
  - **Solves problems of traditional file environment**
    - Controls redundancy
    - Eliminates inconsistency
    - Uncouples programs and data
    - Enables organization to central manage data and data security

HUMAN RESOURCES DATABASE WITH MULTIPLE VIEWS

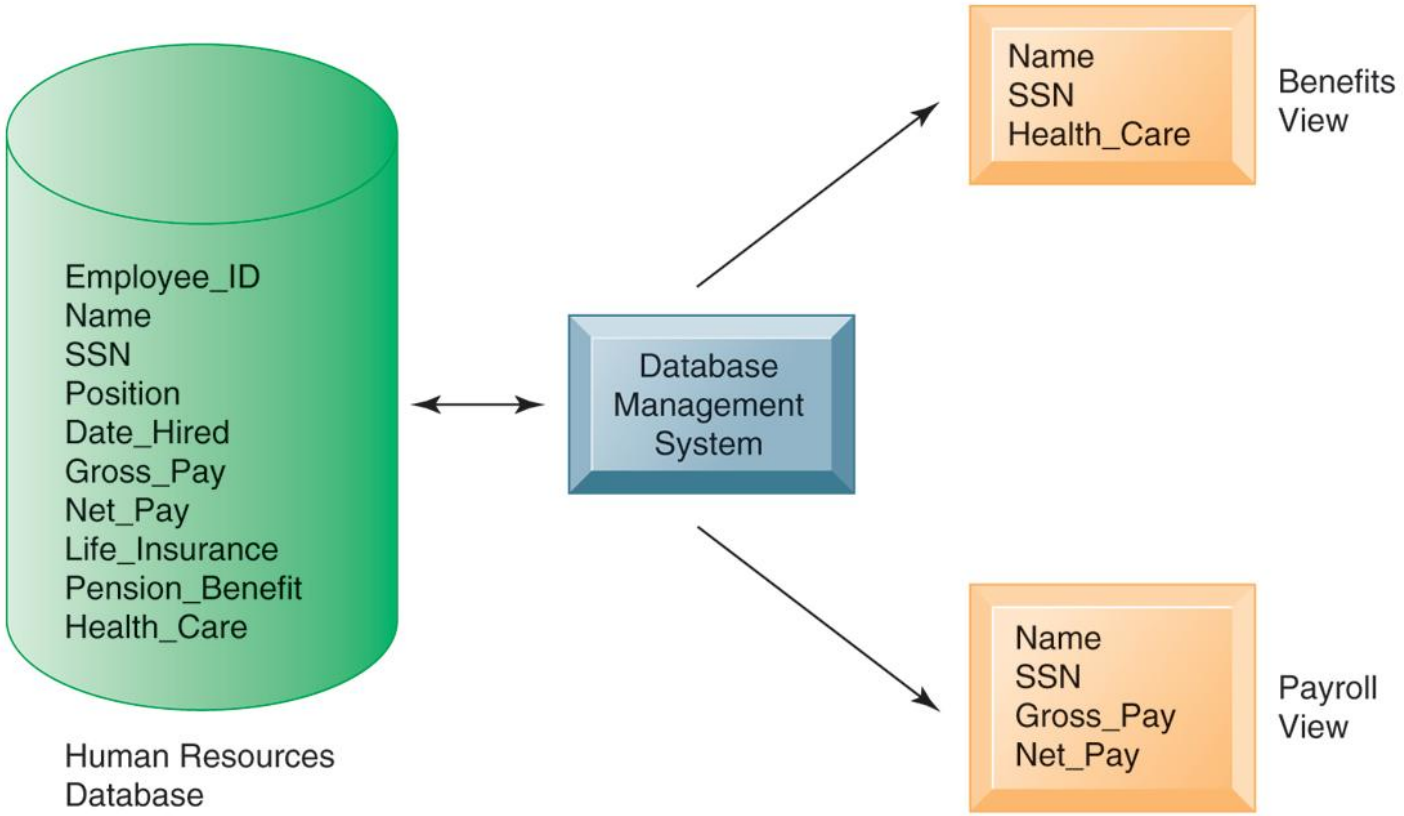


FIGURE 6-3 A single human resources database provides many different views of data, depending on the information requirements of the user. Illustrated here are two possible views, one of interest to a benefits specialist and one of interest to a member of the company's payroll department.

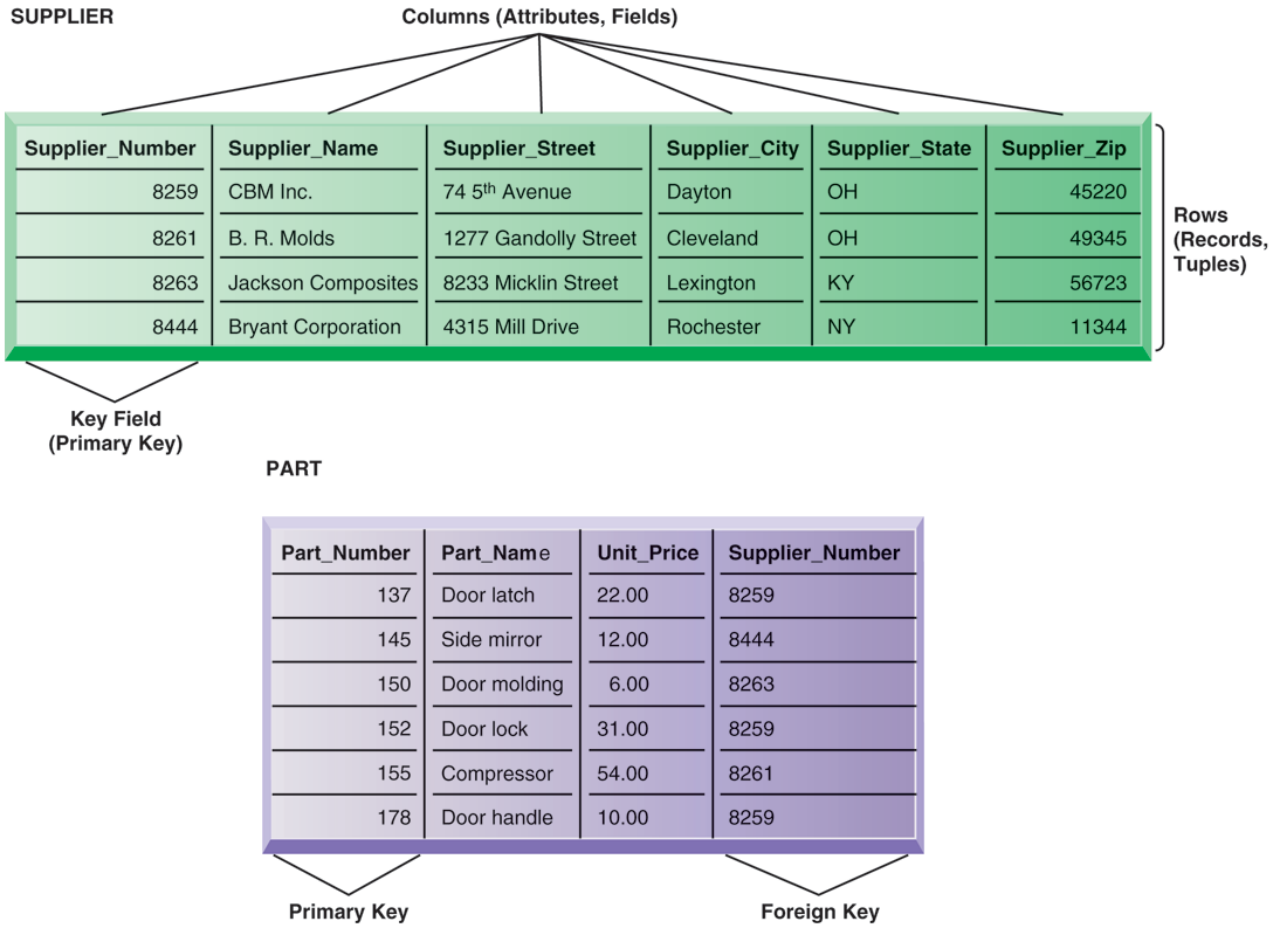
Capabilities of Database Management Systems (DBMSs)

- **Relational DBMS**
  - Represent data as two-dimensional tables
  - Each table contains data on entity and attributes
- **Table: grid of columns and rows**
  - Rows (tuples): Records for different entities
  - Fields (columns): Represents attribute for entity
  - Key field: Field used to uniquely identify each record
  - Primary key: Field in table used for key fields
  - Foreign key: Primary key used in second table as look-up field to identify records from original table

Relational Database Tables

A relational database organizes data in the form of two-dimensional tables. Illustrated here are tables for the entities SUPPLIER and PART showing how they represent each entity and its attributes. Supplier Number is a primary key for the SUPPLIER table and a foreign key for the PART table.

FIGURE 6-4



- **Operations of a Relational DBMS**
  - **Three basic operations used to develop useful sets of data**
    - **SELECT:** Creates subset of data of all records that meet stated criteria
    - **JOIN:** Combines relational tables to provide user with more information than available in individual tables
    - **PROJECT:** Creates subset of columns in table, creating tables with only the information specified

THE THREE BASIC OPERATIONS OF A RELATIONAL DBMS

PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Select Part_Number = 137 or 150

SUPPLIER

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Components	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Join by Supplier_Number

Part_Number	Part_Name	Supplier_Number	Supplier_Name
137	Door latch	8259	CBM Inc.
150	Door molding	8263	Jackson Components

Project selected columns

FIGURE 6-5 The select, join, and project operations enable data from two different tables to be combined and only selected attributes to be displayed.

Capabilities of Database Management Systems (DBMSs)

- **Non-relational databases: "NoSQL"**
  - More flexible data model
  - Data sets stored across distributed machines
  - Easier to scale
  - Handle large volumes of unstructured and structured data (Web, social media, graphics)
- **Databases in the cloud**
  - Typically, less functionality than on-premises DBs
  - Amazon Relational Database Service, Microsoft SQL Azure
  - Private clouds

Capabilities of Database Management Systems (DBMSs)

- **Capabilities of database management systems**
  - **Data definition capability: Specifies structure of database content, used to create tables and define characteristics of fields**
  - **Data dictionary: Automated or manual file storing definitions of data elements and their characteristics**
  - **Data manipulation language: Used to add, change, delete, retrieve data from database**
    - Structured Query Language (SQL)
    - Microsoft Access user tools for generating SQL
  - **Many DBMS have report generation capabilities for creating polished reports (Crystal Reports)**

MICROSOFT ACCESS DATA DICTIONARY FEATURES

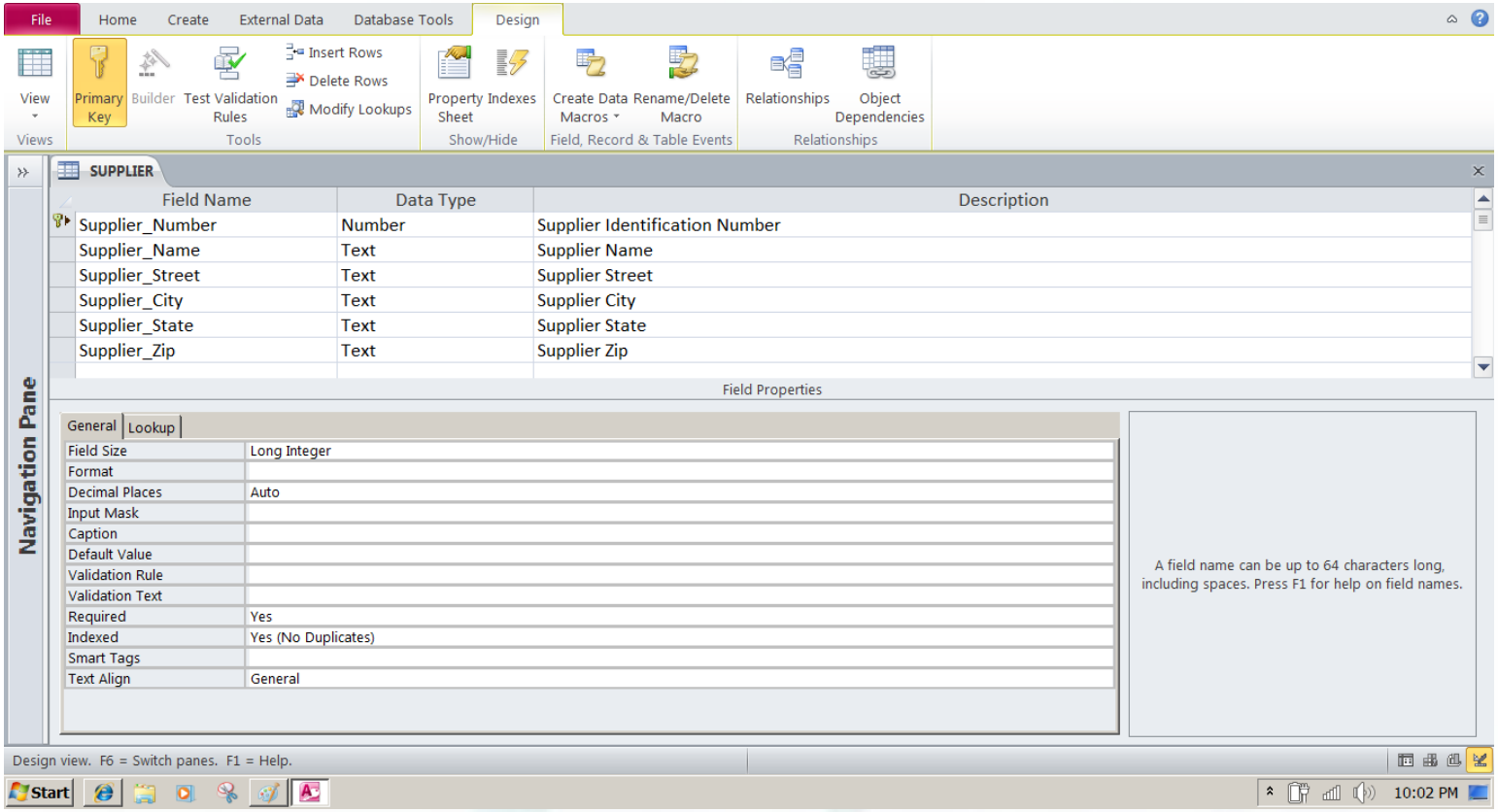


FIGURE 6-6 Microsoft Access has a rudimentary data dictionary capability that displays information about the size, format, and other characteristics of each field in a database. Displayed here is the information maintained in the SUPPLIER table. The small key icon to the left of Supplier_Number indicates that it is a key field.

EXAMPLE OF AN SQL QUERY

```
SELECT PART.Part_Number, PART.Part_Name, SUPPLIER.Supplier_Number,
SUPPLIER.Supplier_Name
FROM PART, SUPPLIER
WHERE PART.Supplier_Number = SUPPLIER.Supplier_Number AND
Part_Number = 137 OR Part_Number = 150;
```

FIGURE 6-7 Illustrated here are the SQL statements for a query to select suppliers for parts 137 or 150. They produce a list with the same results as Figure 6-5.

AN ACCESS QUERY

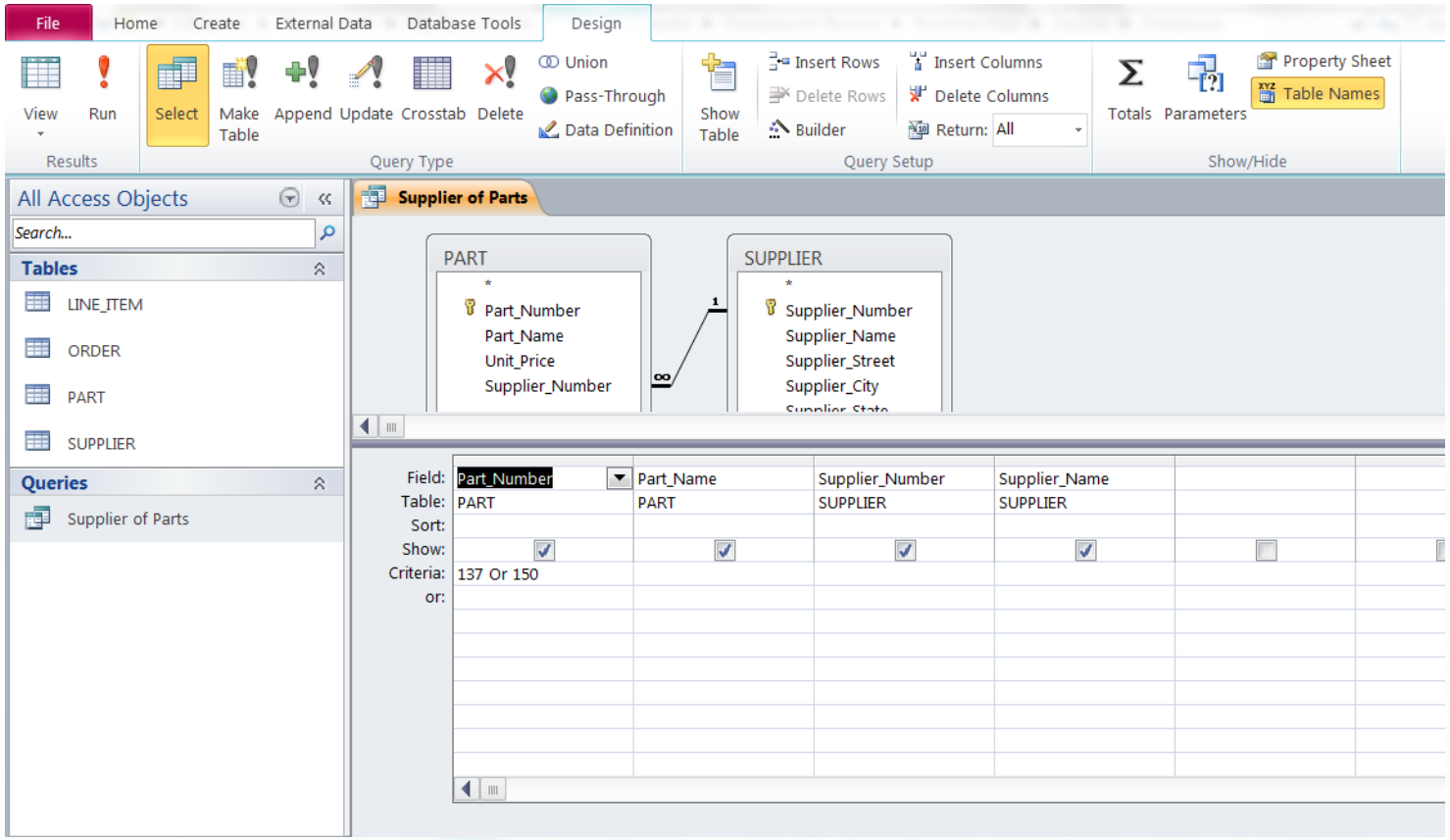


FIGURE 6-8 Illustrated here is how the query in Figure 6-7 would be constructed using Microsoft Access query building tools. It shows the tables, fields, and selection criteria used for the query.

Capabilities of Database Management Systems (DBMSs)

- **Designing Databases**
  - Conceptual (logical) design: abstract model from business perspective
  - Physical design: How database is arranged on direct-access storage devices
- **Design process identifies:**
  - Relationships among data elements, redundant database elements
  - Most efficient way to group data elements to meet business requirements, needs of application programs
- **Normalization**
  - Streamlining complex groupings of data to minimize redundant data elements and awkward (unsuitable/unapproperiate) many-to-many relationships

AN UNNORMALIZED RELATION FOR ORDER

FIGURE 6-9 An unnormalized relation contains repeating groups. For example, there can be many parts and suppliers for each order. There is only a one-to-one correspondence between Order_Number and Order_Date.

NORMALIZED TABLES CREATED FROM ORDER

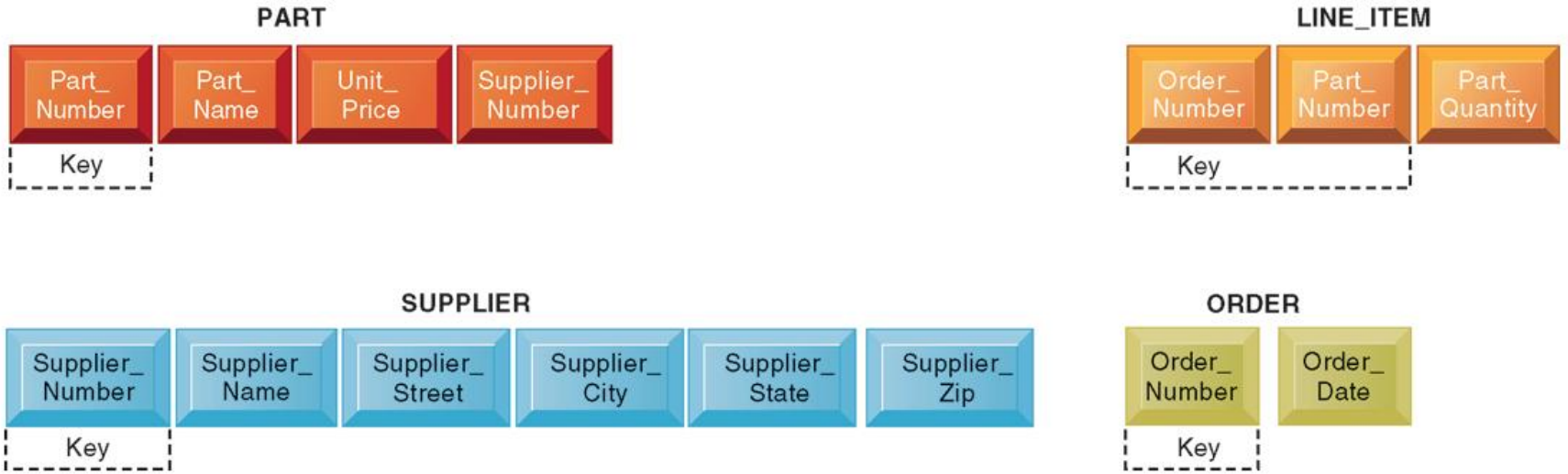


FIGURE 6-10 After normalization, the original relation ORDER has been broken down into four smaller relations. The relation ORDER is left with only two attributes and the relation LINE_ITEM has a combined, or concatenated, key consisting of Order_Number and Part_Number.

Capabilities of Database Management Systems (DBMSs)

- **Referential integrity rules**
  - Used by RDMS to ensure relationships between tables remain consistent
- **Entity-relationship diagram**
  - Used by database designers to document the data model
  - Illustrates relationships between entities
- **Caution: If a business doesn't get data model right, system won't be able to serve business well**

AN ENTITY-RELATIONSHIP DIAGRAM

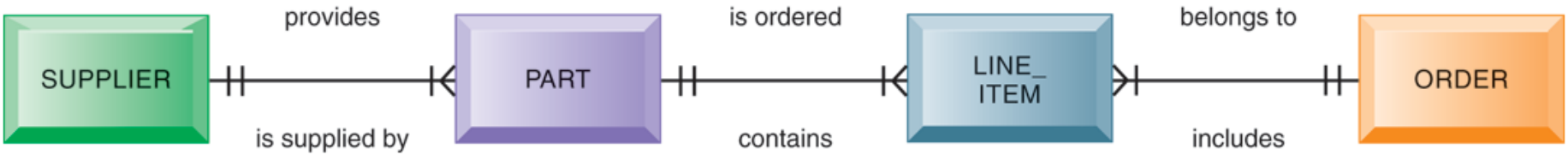


FIGURE 6-11 This diagram shows the relationships between the entities SUPPLIER, PART, LINE_ITEM, and ORDER that might be used to model the database in Figure 6-10.

Tools for Improving Business Performance and Decision Making

- **Big data**
  - **Massive (huge) sets of unstructured/semi-structured data from Web traffic, social media, sensors, and so on**
  - **Petabytes, exabytes of data**
  - **The petabyte is a multiple of the unit byte for digital information. The prefix peta indicates the fifth power of 1000 and means 10^{15} in the International System of Units (SI), and therefore 1 petabyte is one quadrillion (short scale) bytes, or 1 billiard (long scale) bytes. The unit symbol for the petabyte is PB. 1 PB = 1000000000000000B = 10^{15}bytes = 1000 terabytes.**
    - **Exabyte (1 000 000 000 000 000 000 bytes=10^{18} bytes)**
    - **Zettabyte (1 000 000 000 000 000 000 000 bytes=10^{21} bytes)**
    - **Yottabyte (1 000 000 000 000 000 000 000 000 bytes=10^{24} bytes)**
    - **Xenottabyte (1 000 000 000 000 000 000 000 000 000 bytes=10^{27} bytes)**
    - **Shilentnobyte (1 000 000 000 000 000 000 000 000 000 000 bytes=10^{30} bytes)**
    - **Domegemegrottebyte (1 000 000 000 000 000 000 000 000 000 000 000 bytes=10^{33} bytes)**
    - Volumes too great for typical DBMS
  - **Can reveal more patterns and anomalies**

Tools for Improving Business Performance and Decision Making

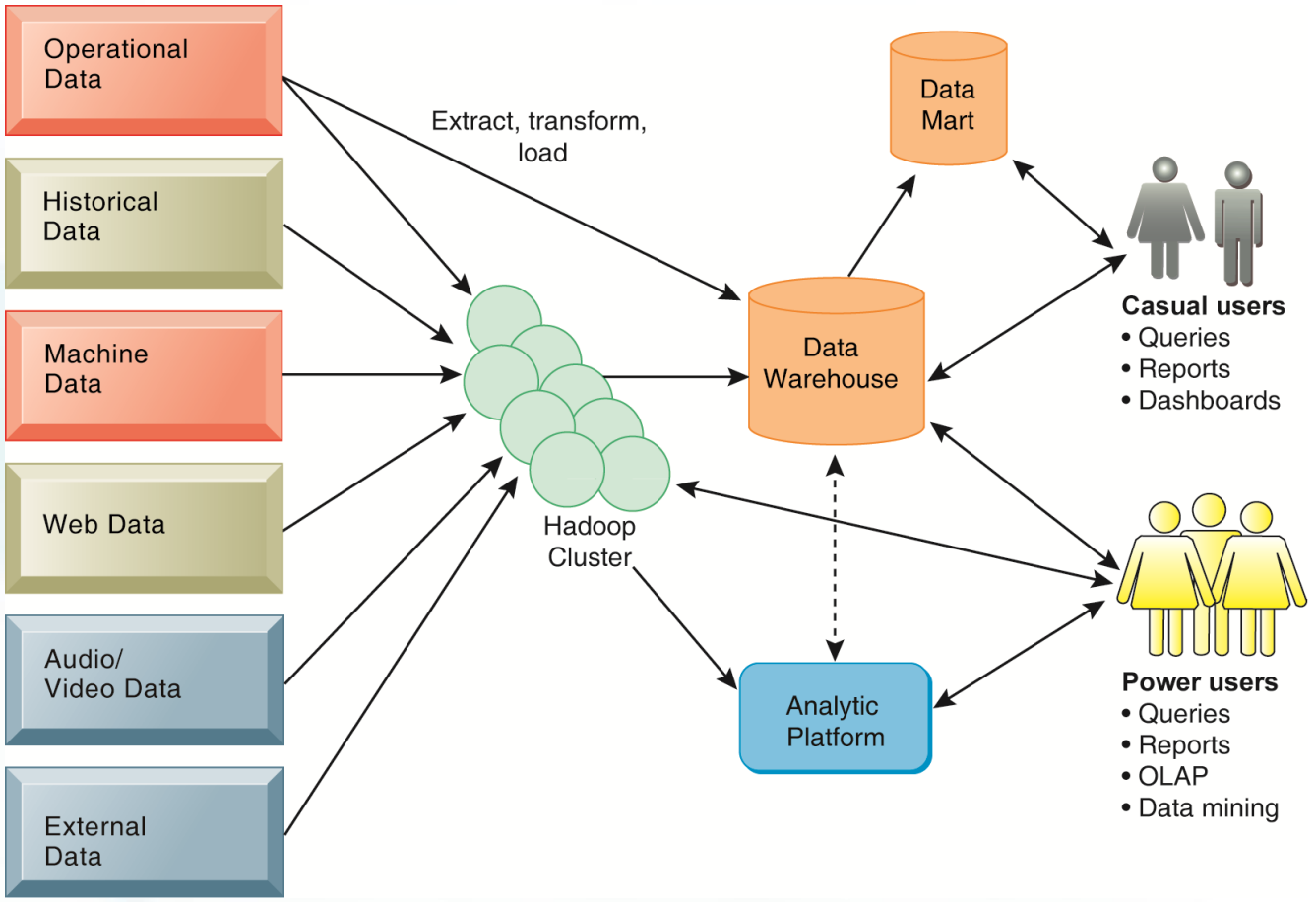
- **Business intelligence infrastructure**
  - Today includes an array of tools for separate systems, and big data
- **Contemporary tools:**
  - Data warehouses
  - Data marts
  - Hadoop
  - In-memory computing
  - Analytical platforms

- **Data warehouse:**
  - Stores current and historical data from many core operational transaction systems
  - Consolidates and standardizes information for use across enterprise, but data cannot be altered
  - Provides analysis and reporting tools
- **Data marts:**
  - Subset of data warehouse
  - Summarized or focused portion of data for use by specific population of users
  - Typically focuses on single subject or line of business

CONTEMPORARY BUSINESS INTELLIGENCE INFRASTRUCTURE

A contemporary business intelligence infrastructure features capabilities and tools to manage and analyze large quantities and different types of data from multiple sources. Easy-to-use query and reporting tools for casual business users and more sophisticated analytical toolsets for power users are included.

FIGURE 6-12



- **Hadoop**
  - **Enables distributed parallel processing of big data across inexpensive computers**
  - **Key services**
    - Hadoop Distributed File System (HDFS): data storage
    - MapReduce: breaks data into clusters for work
    - Hbase: NoSQL database
  - **Used by Facebook, Yahoo, NextBio**

- **In-memory computing**
  - Used in big data analysis
  - Uses computers main memory (RAM) for data storage to avoid delays in retrieving data from disk storage
  - Can reduce hours/days of processing to seconds
  - Requires optimized hardware
- **Analytic platforms**
  - High-speed platforms using both relational and non-relational tools optimized for large datasets

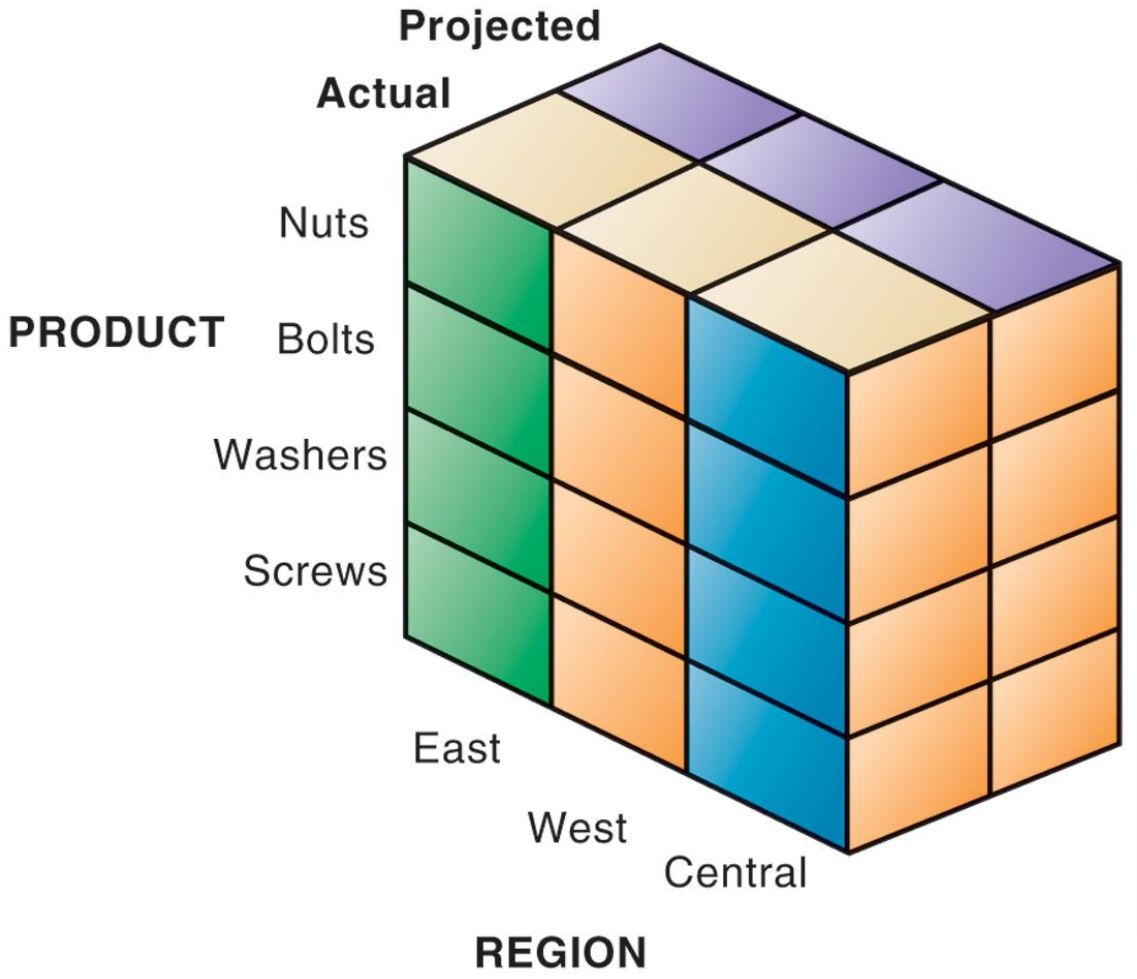
- **Analytical tools: Relationships, patterns, trends**
  - **Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions**
    - Multidimensional data analysis (OLAP)
    - Data mining
    - Text mining
    - Web mining

- **Online analytical processing (OLAP)**
  - **Supports multidimensional data analysis**
    - Viewing data using multiple dimensions
    - Each aspect of information (product, pricing, cost, region, time period) is different dimension
    - Example: How many washers sold in the East in June compared with other regions?
  - **OLAP enables rapid, online answers to ad hoc queries**

MULTIDIMENSIONAL DATA MODEL

The view that is showing is product versus region. If you rotate the cube 90 degrees, the face that will show product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. Other views are possible.

FIGURE 6-13



- **Data mining:**
  - **Finds hidden patterns, relationships in datasets**
    - Example: customer buying patterns
  - **Infers rules to predict future behavior**
  - **Types of information obtainable from data mining:**
    - Associations
    - Sequences
    - Classification
    - Clustering
    - Forecasting

- **Text mining**
  - **Extracts key elements from large unstructured data sets**
    - Stored e-mails
    - Call center transcripts
    - Legal cases
    - Patent descriptions
    - Service reports, and so on
  - **Sentiment analysis software**
    - Mines e-mails, blogs, social media to detect opinions

- **Web mining**
  - **Discovery and analysis of useful patterns and information from Web**
    - Understand customer behavior
    - Evaluate effectiveness of Web site, and so on
  - **Web content mining**
    - Mines content of Web pages
  - **Web structure mining**
    - Analyzes links to and from Web page
  - **Web usage mining**
    - Mines user interaction data recorded by Web server

- **Databases and the Web**
  - **Many companies use Web to make some internal databases available to customers or partners**
  - **Typical configuration includes:**
    - Web server
    - Application server/middleware/CGI scripts
    - Database server (hosting DBMS)
  - **Advantages of using Web for database access:**
    - Ease of use of browser software
    - Web interface requires few or no changes to database
    - Inexpensive to add Web interface to system

LINKING INTERNAL DATABASES TO THE WEB

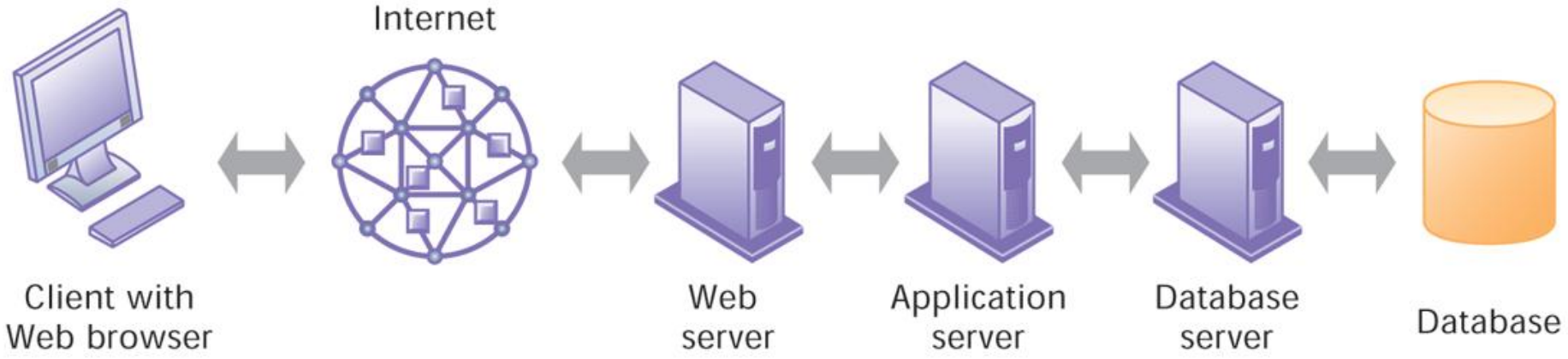


FIGURE 6-14 Users access an organization's internal database through the Web using their desktop PCs and Web browser software.

- **Establishing an information policy**
  - **Firm's rules, procedures, roles for sharing, managing, standardizing data**
  - **Data administration**
    - Establishes policies and procedures to manage data
  - **Data governance**
    - Deals with policies and processes for managing availability, usability, integrity, and security of data, especially regarding government regulations
  - **Database administration**
    - Creating and maintaining database

- **Ensuring data quality**
  - **More than 25 percent of critical data in Fortune 1000 company databases are inaccurate or incomplete**
    - Redundant data
    - Inconsistent data
    - Faulty input
  - **Before new database in place, need to:**
    - Identify and correct faulty data
    - Establish better routines for editing data once database in operation

- **Data quality audit:**
  - **Structured survey of the accuracy and level of completeness of the data in an information system**
    - Survey samples from data files, or
    - Survey end users for perceptions of quality
- **Data cleansing**
  - **Software to detect and correct data that are incorrect, incomplete, improperly formatted, or redundant**
  - **Enforces consistency among different sets of data from separate information systems**

Interactive Session: Management

American Water Keeps Data Flowing

Read the Interactive Session and discuss the following questions

- **Discuss the role of information policy, data administration, and efforts to ensure data quality in improving data management at American Water.**
- **Describe roles played by information systems specialists and end users in American Water's systems transformation project.**
- **Why was the participation of business users so important? If they didn't play this role, what would have happened?**

Interactive Session: Management

American Water Keeps Data Flowing (cont.)

Read the Interactive Session and discuss the following questions

- **How did implementing a data warehouse help American Water move toward a more centralized organization?**
- **Give some examples of problems that would have occurred at American Water if its data were not "clean"?**
- **How did American Water's data warehouse improve operations and management decision making?**